Topic Identification of Arabic Texts based on Statistical Techniques.

L. Fodil, S. Ouamour
TDE, Laboratory of Speech Communication and Signal Processing.

Abstract— One of the main factors that characterize a text is its content. Nowadays, the number of documents scattered online by private and public sectors are in the orders of millions. The rapid growth in the number of documents necessitates the use of automatic text classification. While a lot of effort has been put into manifold languages, minimal experimentation has been done with Arabic. Arabic language is highly inflectional and derivational language which makes text mining a challenging task.

In this paper, we propose two statistical approaches for topic identification. In the first approach we have developed two techniques ACM (Automatic Classification Method) and SACM (Semi-Automatic Classification Method) for the keywords extraction. In the second approach, we have used Centroid Classifier Models to classify the text documents by employing several distances (Euclidean, Manhattan, chebychev, etc.). The tests of evaluation are conducted on an Arabic textual corpus containing 5 different topics: Economics, Politics, Sport, Medicine and Religion. Results show the efficiency of the proposed approaches on topic identification.

Keywords— Arabic Language, Topic Identification, Text Categorization.

I. Introduction

Documents represent the principal repositories of knowledge and the most effective way to illustrate ideas, thoughts, and expertise (Khorsheed, 2013). Nowadays, the volume of document available on the World Wide Web and databases is increasing. The process of discovering and producing the hidden and useful information, embedded inside these documents, manually by domain experts is extremely hard and time consuming. This is since the numbers of online textual data are numerous and these data have large dimensionality. Therefore, using intelligent way such, as through text classification, to discover the benefit of the knowledge they contain